Case Study

Predict High-Risk Donors for Hepatitis C:
Predict high risk donors using the 'K-Nearest Neighbors' model

KNN 모델을 활용한 HCV 고위험 헌혈자 선별

Hyun Jin (Austin) Kang, Yoojin (Audrey) Jung (Data Analyst)

Table of Contents

1. Background
2. Objective
   a. Identify which variables really impact blood donation rate
3. Plan (Prepare and Collect Data)
   a. Data collection * pre-processing
4. Analyze (Visualization)
   a. Data Visualization
5. Construct model (Design and Test models)
   a. Find Feature Importances: Random Forest Model
   b. Classifying Target: K-Nearest Neighbors(KNN) Model
   c. Summary of model results
6. Execute
   a. Conclusion (Recommendations)

Background & Objective

Background

- Hepatitis C (HCV) is known to be a 'transfusion-transmissible infections(TTI)'

- World Health Organization(WHO) recommends all blood donations should be screened for HCV*

- Predicting donors with a high-risk of Hepatitis C (before testing HCV) will help blood centers to pre-screen and monitor high-risk donors, enhancing blood safety.

WHO, Blood safety and availability

Objective

- Predict blood donor with a high risk of Hepatitis C

- Identify which variables have an impact on classifying Hepatitis C patient

Pace : Plan Stage

# Plan (Data Collection & Preprocessing)

- Data Collection
  - Data from UCI Machine Learning Repository*
  - Data has 615 people data including 7 blood test results
    - Blood Test Items: 'ALB', 'ALP', 'ALT', 'AST', 'BIL', 'CHE', 'CHOL', 'CREA', 'GGT', 'PROT'
  - Target attribute has 4 types of categories: blood donors, hepatitis C, fibrosis, and Cirrhosis
- Pre-processing
  - Replaced 31 missing values with the group mean of each category.
  - Merged all Hepatitis C patient's data into one category; 'patient'
  - Excluded 'suspect blood donor' data from modeling for final test (prevent data leakage)

* Lichtinghagen,Ralf, Klawonn,Frank, and Hoffmann,Georg. (2020). HCV data. UCI Machine Learning Repository. https://doi.org/10.24432/C5D612.

pAce: Analyze Stage

Data Distribution

- **Identified outliers across blood test items**
  → Removed outliers(using Z score) on selected items considering the characteristic of the KNN model
  → Scaled data before training

Patient Data Distribution

Non-Patient Data Distribution
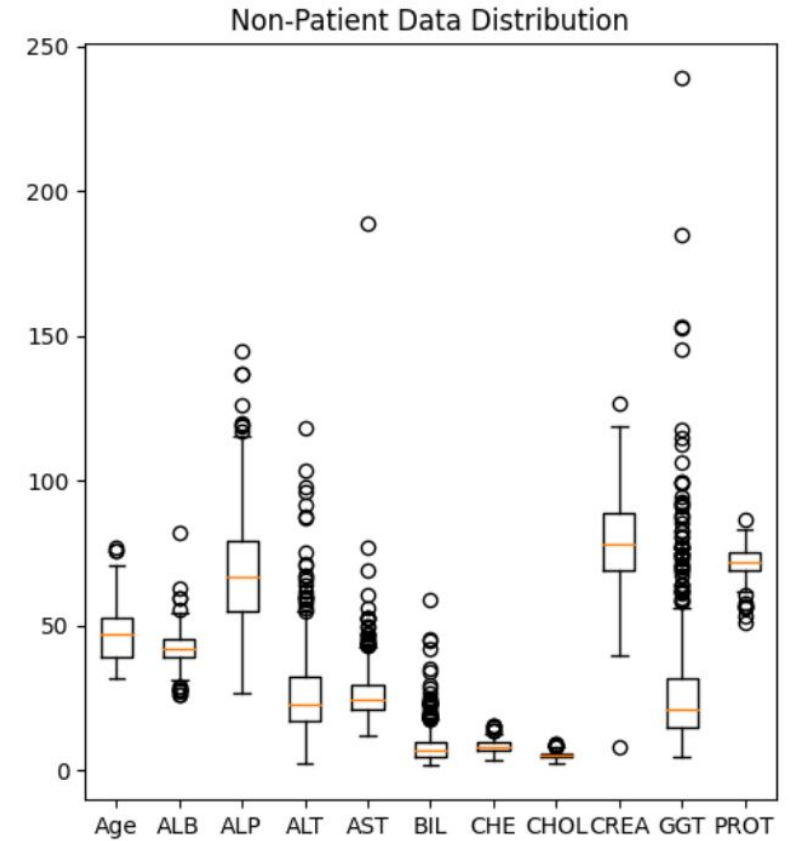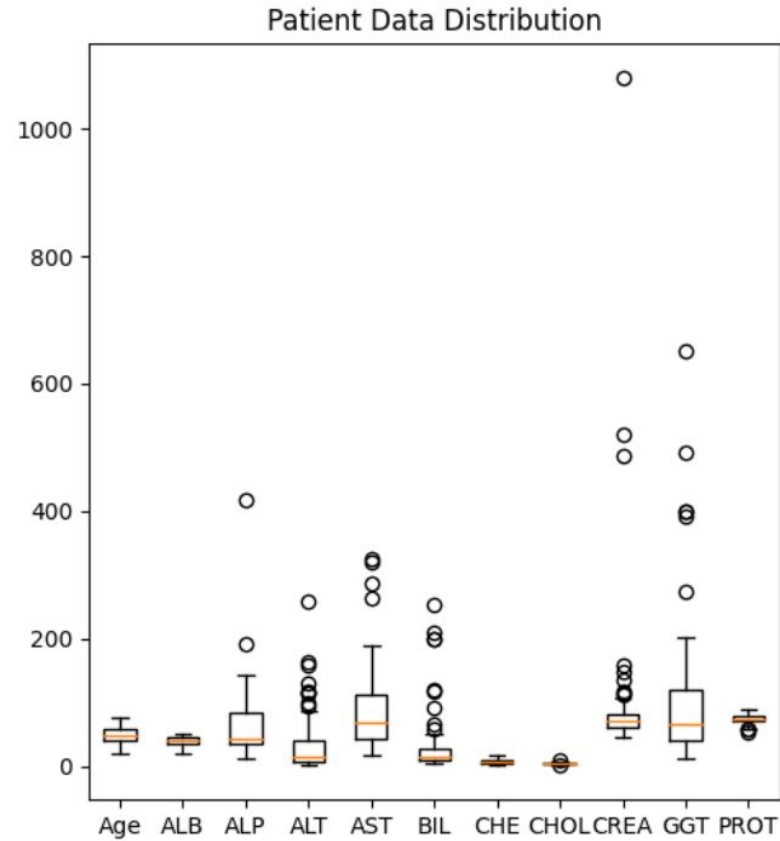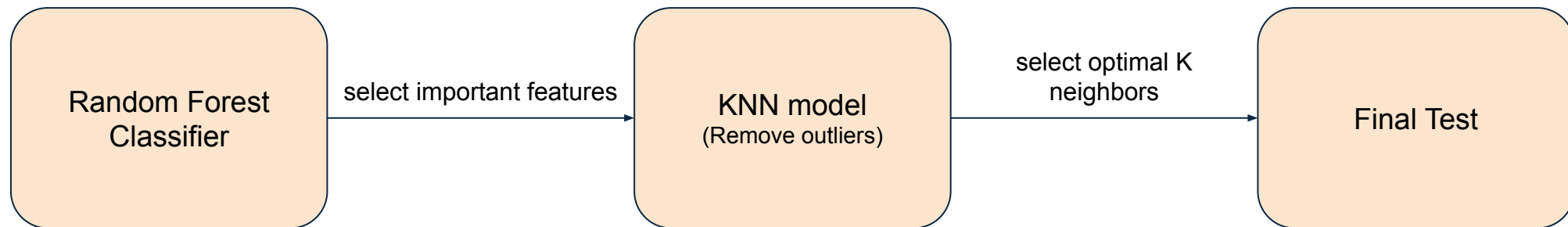
Relationship between features

- Identified there are some correlations between variables.
  - PROT - ALB
  - GGT - AST
  - CHE - CHOL
- To prevent dimensionality reduction and overfitting, used a random forest model to find important features
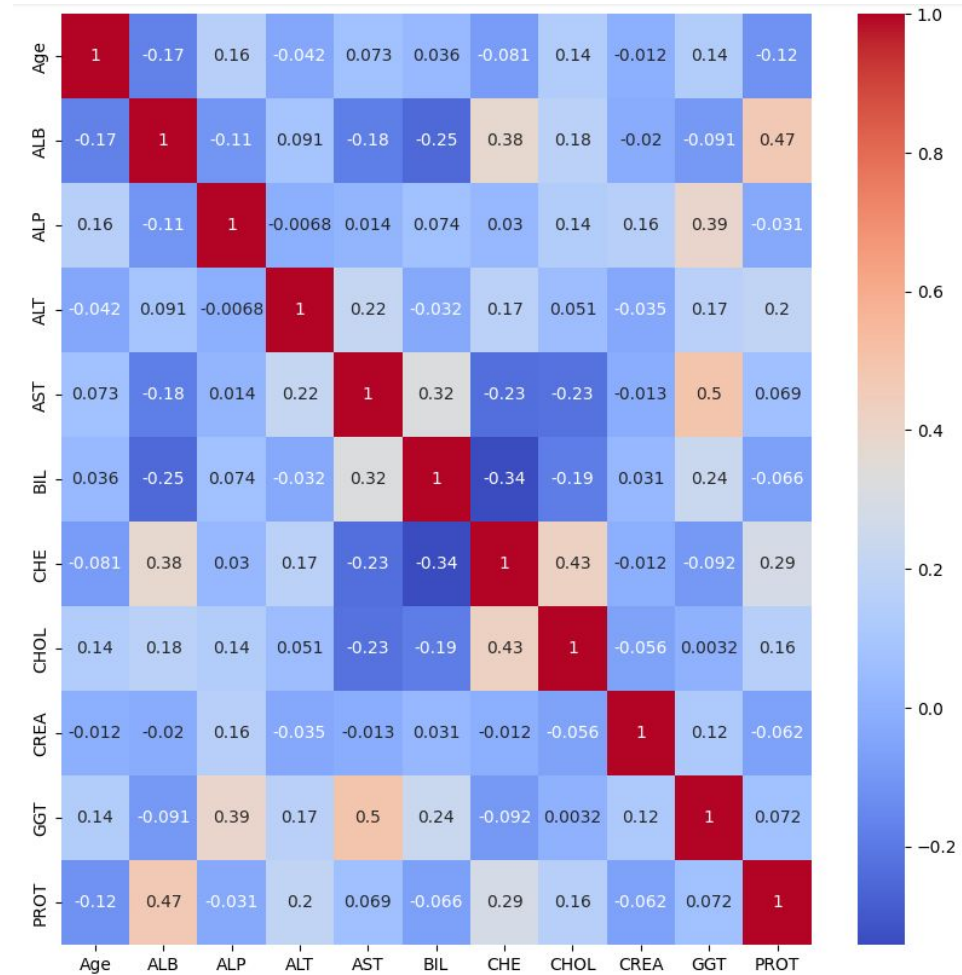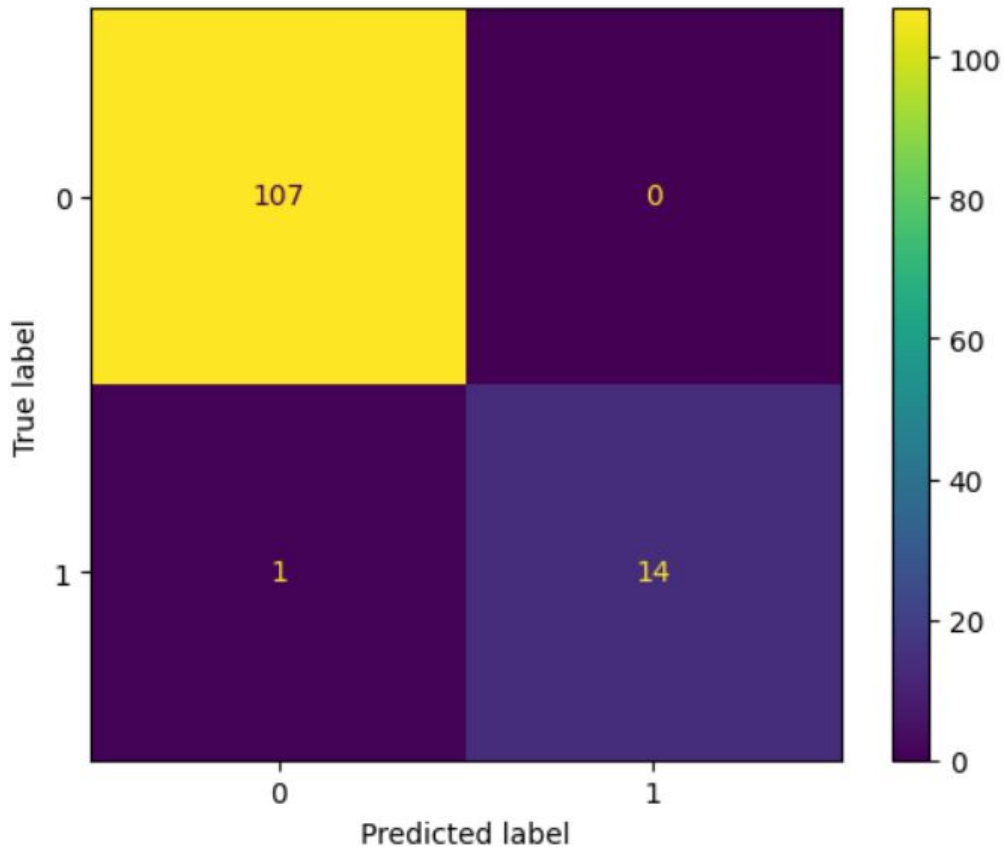
paCe: Construct Stage

Prepare Modeling

- The goal of this study is to classify(and predict) Hepatitis C patient
- To select important features, used a Random Forest Classifier, then built a K-NN model using selected features
- The KNN model is a simple algorithm with fewer parameters to adjust. The model is capable of classifying with small datasets and high dimensionality.
- Due to the class imbalance, used undersampling and SMOTE method (80% non-patient : 20% patient)
  a. selected F1 score as the main evaluation score
- Used 'suspect blood donor' category data to predict if the donor has a high-risk of hepatitis C

```
Random Forest          select important features     KNN model         select optimal K          Final Test
Classifier          ------------------------>    (Remove outliers)  --neighbors-->
```

Model Results(Random Forest Classifier)

	Feature	Importance
0	AST	0.387839
1	GGT	0.175390
2	BIL	0.123582
3	ALP	0.121319
4	ALT	0.061418

- From the Random Forest model, able to extract top important features with high evaluation score
  - Important features: AST, GGT, BIL, ALP, ALT
  - F1 score: 0.96
    Accuracy score: 0.99
    ROC score: 0.96
- Added two additional features(ALB, CHOL), considering current testing items at the blood center

Construct Model

Target (y)   Patient (1, 0) → 20%(1) : 80%(0)

Feature (X)   AST, GGT, BIL, ALP, ALT, ALB, CHOL   *Removed outliers by each target class

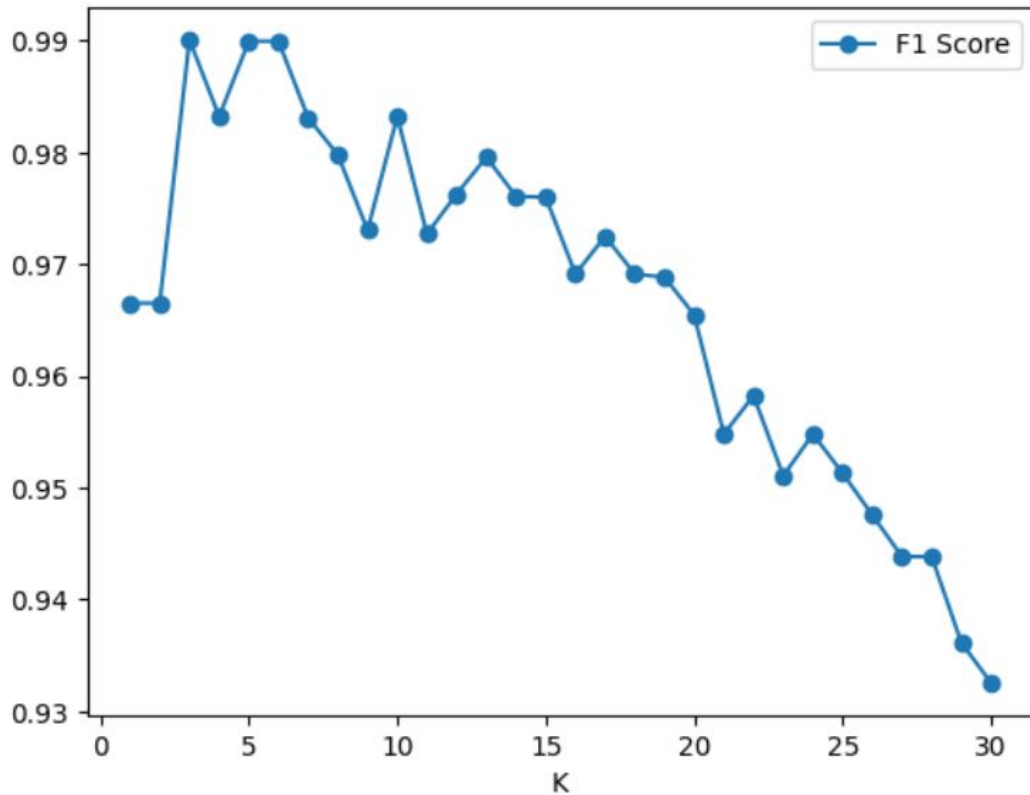**Model**   K-Nearest Neighbor(KNN) Model

Scaling   Downsampling → SMOTE (Upsampling) → Standard Scaler

Evaluation   F1 score, Accuracy score, and ROC score

KNN Model Results

K-Value vs F1 Score

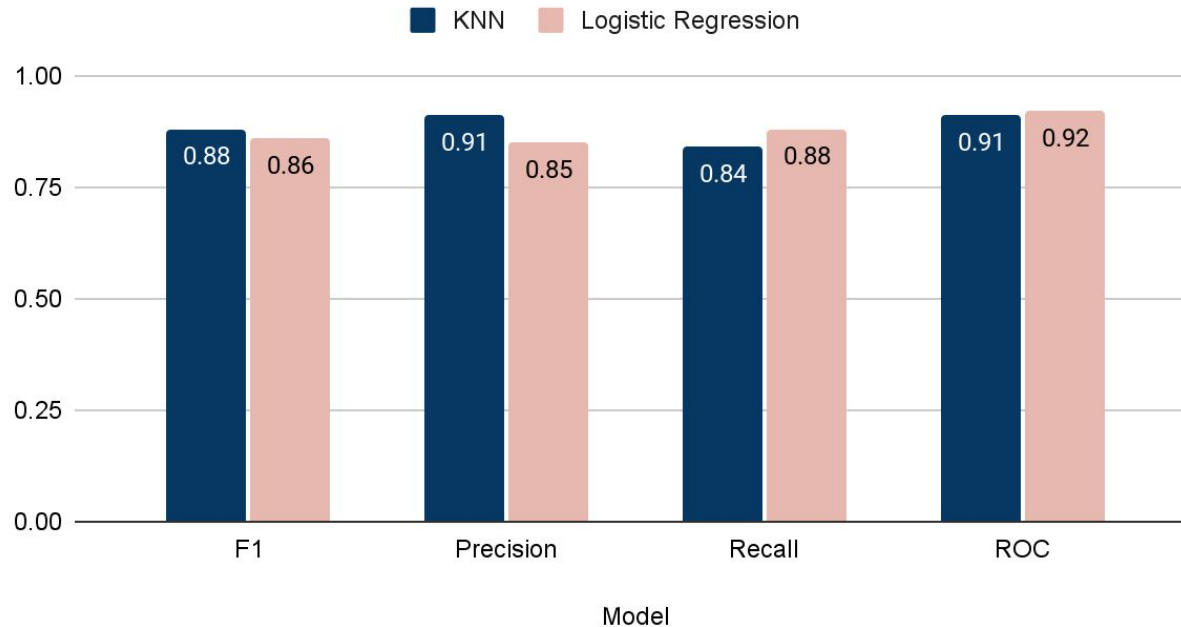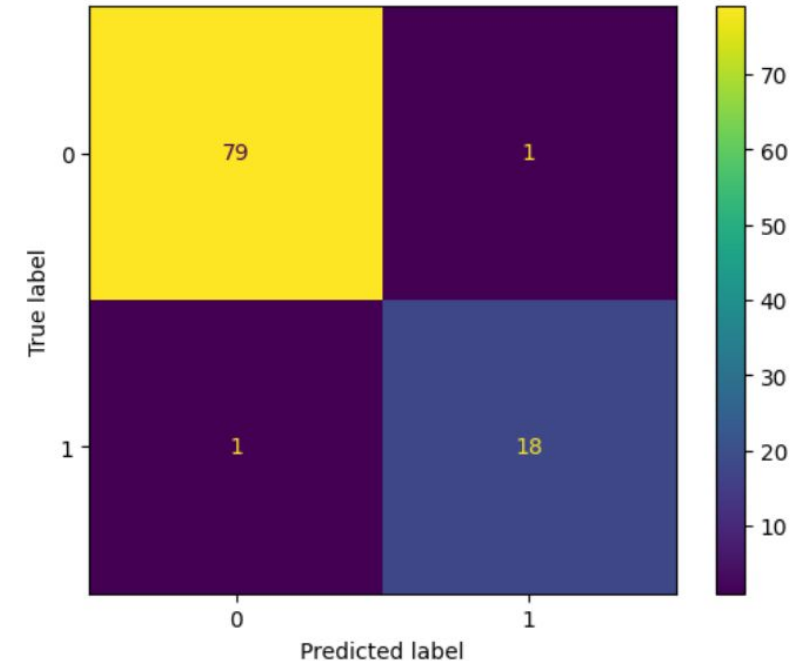|	K	F1 Score
2	3	0.990097
4	5	0.989966
6	7	0.983094
12	13	0.979588
14	15	0.976008

- Found K values of 1, 9, and 3 showed highest F1 score of 0.96
- To prevent 'overfitting' and 'underfitting', chose K values of 3 as an optimal k value

Final Model - KNN model with 3 neighbors

KNN and Logistic Regression

<KNN Model Confusion Matrix>

- In the test data, the KNN model showed a higher score of F1 and Precision

- With an AUC score of 0.91, the KNN model performs better than random guessing 91% of the time.
  → Strong ability to classify between the two classes

pacE: Execute Stage

Predict New Dataset

Suspected Blood Donors

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
534	4	47	m	22.5	124.0	79.5	46.7	2.3	6.83	4.30	170.0	345.6	58.6
535	4	48	m	24.9	116.9	49.2	24.3	4.9	3.44	5.25	29.0	83.0	47.8
536	4	49	m	21.6	42.2	9.5	10.6	2.4	3.75	3.01	64.0	38.9	44.8
537	4	55	m	47.3	106.0	208.8	130.6	0.8	14.80	8.08	76.0	71.6	78.3
538	4	71	m	14.9	69.8	19.7	95.2	9.8	13.30	2.61	9.0	7.6	47.0
539	4	74	m	20.3	84.0	22.8	43.0	5.7	4.91	3.19	52.0	218.3	47.8
540	4	59	f	19.3	208.2	325.3	146.6	6.9	5.33	4.72	32.0	295.6	53.1

Predicted all donors have high risk of HCV

Further investigate these donors' blood samples to prevent any event

```
prediction = optimal_knn_model.predict(new_df_scaled)
print(f'Predicted class for the suspect blood donors : {prediction}')

Predicted class for the suspect blood donors : [1 1 1 1 1 1 1]
```

Conclusion

- Identified AST, GGT, BIL, ALP, and ALT are important features in predicting HCV

- The KNN model achieved an F1-score of 88%, precision of 91%, and AUC-score of 91% on the test set.

- The KNN model is a simple and effective machine learning model to classify the target in a short time with fewer parameters to be tuned.

- Blood centers can utilize this model to classify donors with a high risk of HCV, enhancing blood safety.
    - If the donor is predicted to have a risk of HCV, further investigate his/her previous blood sample and monitor the next blood donation result

Reference

Full Project Code: Github

<Contributions>
- **Yoojin (Audrey) Jung** - Data Analyst (Professional)
  - **(Contribution) EDA (processing & visualization)**
  - (Skills) Data Analytics, Visualization (Tableau), SQL, R, Python
  - (LinkedIn) https://www.linkedin.com/in/yoojin-jung/

- **Hyunjin (Austin) Kang** - Data Scientist (Associate)
  - **(Contribution) Data collection, EDA (feature engineering), construct model**
  - (Interest Area) Public Health
  - (Skills) Machine Learning, Python, SQL
  - (LinkedIn) www.linkedin.com/in/austinkang0702